Data Quality Templates

RPG Templates

6/22/00

Table of Contents

Introduction

Data quality issues concern data producers, data centers, and model and simulation (M&S) developers and users, who all share responsibility for the quality of data used in modeling and simulation. Data quality is established during production. Data producers generate data to meet a specification based on the need to represent some aspect of a defined reality. They conduct tests to validate their production techniques and assessments to verify the quality and accuracy of the resulting data [Rothenberg, 1997].

Data quality is a measure of how well data serve the purpose **for which they were produced**. All data are produced for a purpose, and their quality is directly tied to whether they meet the requirements of that purpose.

Determination of data quality is a data producer function. Data quality assessments are conducted during production against the producer's specifications. Data quality assessment is inherently complex and cannot be represented by a simple numeric value. Rather, it is indicated by the sum of a myriad bits of information about the data that are captured during the data production process and made available to the data user as metadata.

Although data quality addresses the appropriateness of the data for a specified use, there is no reason the data cannot be put to a different use as long as the data user understands the requirements of the original purpose and has some confidence that the data can meet the requirements of the current application. However, even when data are consistent and accurate, they may not be suitable for use in a specific model or appropriate for a specific application. They may be incompatible with other data being used in the simulation, they may be based on assumptions inconsistent with simulation specifications, or they may represent a level of fidelity that is inappropriate for the application.

The results of the data quality assessment are provided to the data users who must determine the appropriateness of the data for their particular application. Although the quality of the data is determined by the data producer, only the user of the data can determine whether the data are of the appropriate quality for the new intended purpose. The credibility of the application depends on the credibility of the data no less than the credibility of the model or simulation itself.

Data Quality Template Development

Background

In 1998, the DoD VV&A Technical Working Group (TWG) sponsored a tiger team to examine data V&V and its relationship to M&S VV&A. Results of this effort, including a

set of data quality templates, discussed below, and an integrated M&S and data V&V process, are recorded in the tiger team's final report [White Paper, 1998]. An expansion of the integrated data V&V process is provided in the special topic on data V&V.

Template Structure

The data quality templates were developed to

- assist data producers in providing information useful to data users
- guide data users in obtaining the type of producer-generated data quality information needed to support their data selection and V&V activities,

The templates were developed from the data user's perspective and then mapped to metadata templates used by ADS and by the ISO to ensure consistency and completeness.

They define three levels of metadata:

Template Metadata Levels	
Level	**Description**
Database	Information pertaining to all entries in the database (e.g., single shot kill probabilities (SSKPs) for all threat systems in a given scenario);
Data Element	Information pertaining to all entries concerning a specific data element (e.g., SSKP for threat tanks against all systems in a given scenario);
Data Value	Information pertaining to a specific data value (e.g., SSKP for a threat tank against an Apache helicopter in a given scenario).

Because the listing of all possible metadata needed to support data use is extensive, the information fields have been prioritized using the priority designations shown below:

Information Field Priorities	
1	essential information
2	recommended information
3	nice to have information

Template Usage

The fields are intended to be filled using a top-down approach. The data producer should provide information at the database level first, then fill in information at the other two levels by exception.

- Information is included at the data element level only when it differs from or provides additional detail to that provided at the database level
- information is provided at the data value level only when it differs from or provides additional detail to that provided at the two more aggregated levels.

The data user review the metadata in selected fields to determine the data's appropriateness for use. At a minimum, a metadata review should include the items shown below:

Minimum Metadata Review Items
description -- including resolution, meaning, intended uses, etc.
sources and credibility of the sources
quality information (e.g., completeness, accuracy, validity, currency)
results of quality checks, tests and V&V activities conducted by the producer
compliance with standards
usage information pertaining to similar applications
additional metadata fields as needed to address issues of appropriateness and sufficiency for the current application

Once specific data have been selected for use, additional metadata fields can then be selected to support verification and validation activities.

Tiger Team Data Quality (DQ) Templates

In the data quality (DQ) templates that follow, "essential" information (1) is shown in **red**, "necessary" information (2) is shown in **blue**, and "nice to have" information (3) is shown in **green**.

DQ Metadata – Identification Information*		
Priority	**Metadata**	**Definition**
	Identification	**Basic information about the Information Asset**
1	Primary Point of Contact	Identification of and means of communication with person(s) and organization(s) associated with the information asset. (Responsible Party information)
1	Language of Asset	Language(s) used within the asset
3	Information-Asset Type Code (New)	The code for a kind of Information Asset. Enumerations: document, dataset, databank, SME, date warehouse, library, repository, software program, other
1	Abstract	Brief narrative summary of the Information Asset
1	Purpose	Summary of the intentions with which the information asset was developed
3	Supplemental Information	Other descriptive information about the information asset. Examples: logical data model, physical data model, activity/process models, work flow models, other
1	Data Dictionary Name (new)	The name of the collective definition vehicle for data elements of interest
2	Information Asset Environment	Description of the asset in the producer's processing environment, including items such as the name of the software, computer operating system, filename, and the information asset size
1	Identification Citation	Recommended reference to be used for the information asset (Citation Information)
1	Initiative Identification Information	Direction empowering organization to construct the Information Asset
	Status	**State of and maintenance information for the asset**
3	Progress Code	Status of the information asset
3	Maintenance Information	Information categories about the level and frequency of updating Enumerations: 1 - Completed, 2 - In work, 3 - Planned, 4 - Required, 5 - On-going, 6 - Historical archive, 7 - Obsolete.
2	Maintenance & Update Frequency	Frequency with which changes are made tot he information asset after initial asset is complete. Enumerations are: 2 - Daily, 4 - Weekly, 6 - Biennially, 8 - As needed, 10 - None planned
2	Update Level Code	Level at which changes are applied
	Category	**Words or phrases summarizing a subject of the asset**
1	Information-Asset Keyword	Common use word(s) or phrase(s) used to describe the subject
2	Information-Asset Keyword Type Code	Method used to group similar keywords
3	Information-Asset Keyword Thesaurus Name	Name of the formally registered thesaurus or a similar authoritative source of keywords

3	Information-Asset Designation	Designation assigned to an information asset in accordance with Component designation process; source may propose a designation Enumerations: Category I, Category II, Category III.
	Asset Association	**Association of one Information asset to another information asset**
2	Type of Association Code	Justification of the correlation of the two information assets: 1 stereo-mate, 2 larger work citation, 3 cross-reference, 4 source, 5 series, 6 part of a seamless database, 99 other.
1	User Defined Asset Association	Type of association of one asset to another specified by user
1	Association Information	Reference for the associated asset (Citation Information)
	Asset Constraints	**Restrictions on the access and use of the asset**
1	Use Limitations	Any limitation affecting the fitness for use of the information asset. Example: "Not to be used for navigation"
2	Access Constraints	Access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions on obtaining the information asset
2	Use Constraints	Constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on using the information asset
1	Security Classification	Name of the handling restrictions on the information asset. Examples are 'Top Secret', 'Secret', 'Confidential', 'Restricted', 'Unclassified', 'Sensitive'
*** (corresponds to ISO/TC 211 A.4)**		

DQ Metadata - Quality Information*		
Priority	**Metadata**	**Definition**
	Data Quality	**Assessment of quality for either the dataset or an identified group of data**
1	Data Quality Level Code	Specific group of data, if differing from the dataset, to which the quality information applies. Enumerations are: 1 - Dataset Series, 2 - Feature Type, 3 - Attribute Type, 4 - Relationship, 5 - Other Reporting Group, 6 - Feature List
2	Data Quality Assessment	Information on the quality of the quality information level
1	Data Quality Report Type Code	Type of conformance test conducted. Enumerations include: accuracy, currency, completeness, logical consistency, precision, timeliness, clarity of design, flexibility of design, other, (added resolution, portability of data)
1	Qualitative Assessment	Non-quantitative (descriptive) information on the quality of the quality information level
1	Qualitative Narrative Report	Descriptive quality information for the Qualitative Report Type
1	Quantitative Assessment	Quantitative information on the Quality Information Scope's quantitative quality components
1	Quantitative Report	Quantitative information for a component of quality
1	Conformance Specification	Description or name of the document containing the specification against which the quantitative evaluation in conducted
1	Conformance Explanation	Description of the test and methodology yielding the conformance
1	Conformance Code	Results of the test for conformance
2	Data Quality Value Domain	Value resulting from applying the test to the quality information level
2	Data Quality Result	Unit in which the quantitative value is recorded
2	Data Quality Error Statistic Term	Algorithm used to report the data quality value domain
2	Quality Date /Time	Date and time when the quality examination was conducted

* (corresponds to ISO/TC 211 A.6)

DQ Metadata - Lineage Information*		
Priority	**Lineage Metadata**	**Information about the events, parameters, and source data which constructed the asset; information about the responsible parties**
	Usage	**Basic information about specific application(s) for which the asset has been or is being used by different users**
1	Use Contact Information	Information about the asset user.(Responsible Party Information)
1	Use	A brief description of the information asset usage
2	Use Date / Time	Date and Time of asset use
3	User Determined Limitations	Applications for which the asset is not suitable
	Source	
1	Source Description	Description of the information asset, such as events, parameters, and source data, used to create the information asset.
2	Source Date / Time	Date and Time when the source information was collected
2	Source Citation	Document used to authorize production of the source information including specification, business rules, etc. (Citation Information)
	Process	
2	Process Description	Explanation of the events and related parameters or tolerances
2	Process Date / Time	Date and Time when the event was completed
2	Process Contact	Party responsible for the processing step.(Responsible Party Information)
3	Process Rationale (New)	Discussion of the reasons for choosing each process used for the derivation, generation, collection, and transformation of data within the information asset
* (corresponds to ISO/TC 211 A.8)		

DQ Metadata – Distribution Information*		
Priority	**Metadata**	**Information about the distributor of & options for obtaining the information asset**
3	Distributor	Information about the distributor.(Responsible Party Information)
3	Distribution Identifier	Identifier by which the distributor knows the information asset.
3	Distribution Liability	Statement of the liability assumed by the distributor.
2	Custom Ordering Process	Description of custom distribution services available, and the terms and conditions for obtaining those services.
2	Standard Ordering Process	Common ways in which the information asset may be obtained or received, and related instructions.
2	Fees	Fees and terms for retrieving the information asset.
3	Available Date/Time	Date and time when the information asset will be available.
2	Ordering Instructions	General instructions and advice about, and special terms and services provided for, the information asset by the distributor.
2	Turnaround	Typical turnaround time for the filling of an order.
3	Distribution Format Information	Description of the form of the data.
3	Distribution Format Name	Name of the data transfer format.
1	Distribution File Decompression Technique	Recommendations of algorithms or processes that can be applied to read or expand the information asset to which data compression techniques have been applied.
3	Distribution Transfer Size	Size, or estimated size, of the transferred information asset in megabytes.
3	Distribution Format Version Number	Number of the format version.
3	Dial Up Instructions	Information required to access the distribution computer remotely through telephone lines.
3	Distribution Media	Name of the media on which the information asset can be received.
3	Recording Format	Options available or method used to write the information asset to media.
2	Compatibility Information	Description of other limitations or requirements for using the medium, special HW/SW pre- or post-processing, etc.
* (corresponds to ISO/TC 211 A.16)		

DQ Metadata - On-line Information*		
Priority	**Metadata**	**Information about on-line sources from which assets can be obtained**
2	On-line Resource Name	Name of the resource
2	On-line Resource Description	Description of what the resource is/does
3	On-line Resource Linkage	Uniform Resource Locator (URL) to access the resource
4	On-line Resource Function Code	Function performed by the resource
4	On-line Resource Application	Name of the application profile that can be used with the resource
4	On-line Resource Protocol	Connection protocol to be used
* (corresponds to ISO/TC 211 A.28)		

DQ Metadata - Citation Information*		
Priority	**Metadata**	**Description**
1	Info-Asset Title	Name of an Information Asset.
1	Information Asset Alternate Name	Other language name of an Information Asset.
1	Information Asset Short Name	Abbreviated name or acronym of an Information Asset.
1	Citation Responsible Party	Information about the responsible party cited (Responsible Party)
1	Reference Date	Date and time when the asset was or will be published or otherwise made available.
1	Edition	Version of the titled asset.
* (corresponds to ISO/TC 211 A.20)		

DQ Metadata - Responsible Party Information*		
Priority	**Metadata**	**Description**
	Release authority	**Organization/Agency and/or POC authorized to release all or part of the asset for use**
1	Responsible Party Individual Name	Name of the person responsible - SURNAME, given name, title, separated by a delimiter
1	Responsible Party Organization Name	Name of the organization associated with the information asset
2	Responsible Party Organization Identifier	Acronym of the organization associated with the information asset
1	Responsible Party Position Name	Role or position of person responsible
1	Responsible Party Role Code	Role performed by the responsible party. Added enumerations will be: Oversight Authority, Sponsor, Originator, Custodian, Release Authority, Designating Office, Domain Coordinator, Agent, Process Owner, Distributor, Designating Component, other
1	Organization-Individual Role Code (New)	Code indicating the relationship between the individual and organization
	Fields assumed in Name above	
1	Prefix	A title before an individual's name
1	First Name	Given name of the individual
2	Middle Name	Middle name (or initial) of the individual
1	Last Name	Surname of the individual
2	Suffix	A title after an individual's name
* (ISO/TC 211 A.22)		

DQ Metadata – Address Information*		
Priority	**Address (A.24)**	**Description**
1	Postal Address	Address line for the address
1	City	City of the address
1	Administrative Area	State, province, or county of the address
1	Postal Code	ZIP or other postal code of the address
1	Country	Country of the address
1	Voice Telephone	Telephone number by which individual can speak to the organization or individual
3	TDD/TTY Telephone	Telephone number by which hearing-impaired individuals can contact the organization or individual
2	DSN Telephone	Telephone number by which DSN capable users can contact the organization or individual
2	Facsimile Telephone	Telephone number of the facsimile machine of the organization or individual
1	Electronic Mail Address	Address of the electronic mailbox of the organization or individual
2	Address On-line Resource	Address information for on-line resource. (SEE On-line Resource information)
2	Hours of Service	Time period when individual can speak to the organization or individual
3	Contact Instruction	Supplemental instructions on how or when to contact the individual or organization
* (corresponds to ISO/TC 211 A.24)		

References

Rothenberg, Jeff, Rand. "A Discussion of Data Quality for Verification, Validation, and Certification (VV&C) of Data to be Used in Modeling*," Rand Project Memorandum PM-709-DMSO*, Rand, August 1997. This is an essential guide data quality assessment and verification and validation. Includes considerations for metadata used in judging data quality and supporting data verification and validation.

External links in this document:

Authoritative Data Sources (ADS) website: http://165.113.139.100:90/

RPG links in this document:

select menu*: RPG Reference Document*, select item: "DoD Data VV&A Tiger Team White Paper"

select menu*: RPG Special Topics*, select item: "Data V&V for New Simulations"

General References on Data Quality and Data V&V

Air Force Instruction (ARI) 16-1001: *Verification, Validation and Accreditation (VV&A)*, http://xoc.hq.af.mil/kb/docs/vva_afi.html, June 1996.

Annex C, "Data Verification, Validation, and Certification," IEEE 1278.4, *Recommended Practice for Distributed Interactive Simulation -- Verification, Validation, and Accreditation*, 1997.

Army Pamphlet (PAM) 5-11: *Verification, Validation, and Accreditation of Army Models and Simulations*, 15 October 1993 http://www.amso.army.mil/.

DoD 5000.59-P: Modeling and Simulation (M&S) Master Plan, October 1995, http://www.dmso.mil.

DoD 8320.1-M: *Data Administration Procedures*, OASD/C3I, DTIC, Alexandria VA, 1994.

DoD 8320.1-M-1: *Data Element Standardization Procedures*, OASD/C3I, DTIC, Alexandria VA, January, 1993.

DoD 8320.1-M-3: *Data Quality Assurance Procedures*, (draft), OASD/C3I, DTIC, Alexandria, VA, February, 1994.

DoD Guidelines on Data Quality Management, http://ssed1.ncr.disa.mil/srp/datadmn/dqpaper.html.

DoD Modeling and Simulation (M&S) Data Administration Strategic Plan (DASP), DMSO, April 1995.

DoD Modeling and Simulation Office (DMSO) VV&A web-site: http://www.dmso.mil/home/vva/.

SECNAV Instruction 5200.40: *Verification, Validation, and Accreditation (VV&A) of Models and Simulations*, http://navmsmo.hq.navy.mil/policy/directives/ April, 1999.

Model and Simulation Resource Repository (MSRR), http://www.dmso.mil/dmso/portals/msrr.html

Rothenberg Jeff; Stanley, Walter; Hanna, George; Ralston, Mark. *Rand Project Memorandum PM-710-DMSO*, August 1997. This report offers an outstanding theoretical foundation for data verification and validation. It includes a VV&C process model and considerations for structuring individual data V&V efforts with different kinds of data. It also provides a guide for planning both producer and user V&V activities.

Solick, Susan D. "Interaction Between the Data VV&C and M&S V&V Activities of the DIS VV&A Process Model," 15 DIS-033, Fifteenth DIS Workshop, http://www.sisostds.org/doclib/obtain_doc.cfm?record_id=REF_1000101, September 1996. This paper examines data V&V and Certification as it evolved within the 9-step DIS VV&A process model and discusses the interdependence of data V&V and M&S V&V activities.

Standards for the Interoperability of Distributed Simulations (DIS) Workshop, VV&A Subgroup of the Exercise Management and Feedback (EMF) Forum, http://www.sisostds.org/doclib.

§ § § § § § §